Appendix A

Wikipedia, The Free Encyclopedia,

http://en.wikipedia.org/wiki/Protein_conformation

(visited on 3/20/08)

Protein structure

From Wikipedia, the free encyclopedia

Proteins are an important class of biological macromolecules present in all biological organisms, made up of such elements as carbon, hydrogen, nitrogen, phosphorus, oxygen, and sulfur. All proteins are polymers of amino acids. The polymers, also known as polypeptides consist of a sequence of 20 different L-α-amino acids, also referred to as residues. For chains under 40 residues the term peptide is frequently used instead of protein. To be able to perform their biological function, proteins fold into one, or more, specific spatial conformations, driven by a number of noncovalent interactions such as hydrogen bonding, ionic interactions, Van der Waals' forces and hydrophobic packing. In order to understand the functions of proteins at a molecular level, it is often necessary to determine the three dimensional structure of proteins. This is the topic of the scientific field of structural biology, that employs techniques such as X-ray crystallography or NMR spectroscopy, to determine the structure of proteins.

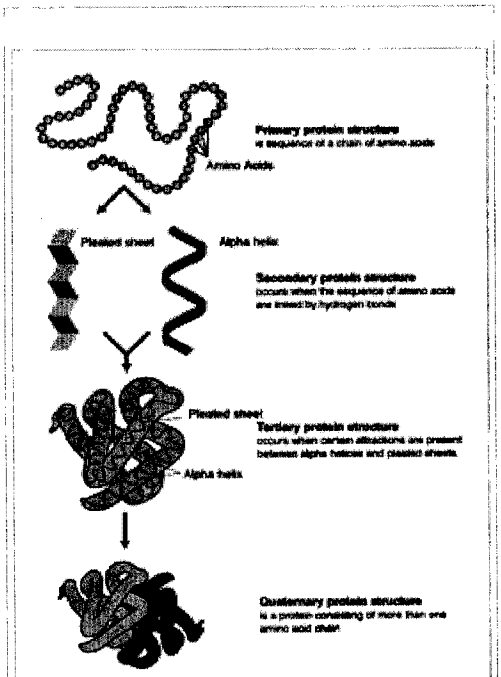
A number of residues are necessary to perform a particular biochemical function, and around 40-50 residues appears to be the lower limit for a functional domain size. Protein sizes range from this lower limit to several thousand residues in multi-functional or structural proteins. However, the current estimate for the average protein length is around 300 residues. Very large aggregates can be formed from protein subunits, for example many thousand actin molecules assemble into a collagen filament.

Contents

Levels of protein structure

Biochemistry refers to four distinct aspects of a protein's structure:

- **Primary structure** - the amino acid sequence of the peptide chains.
- **Secondary structure** - highly regular sub-structures (*alpha helix* and *strands of beta sheet*) which are locally defined, meaning that there can be many different secondary motifs present in one single protein molecule.
- **Tertiary structure** - Three-dimensional structure of a single protein molecule; a spatial arrangement of the secondary structures.
- **Quaternary structure** - complex of several protein molecules or polypeptide chains, usually called protein subunits in this context, which function as part of the larger assembly or protein complex.



Protein structure, from primary to quaternary structure.

In addition to these levels of structure, a protein may shift between several similar structures in performing its biological function. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as chemical conformation, and transitions between them are called conformational changes.

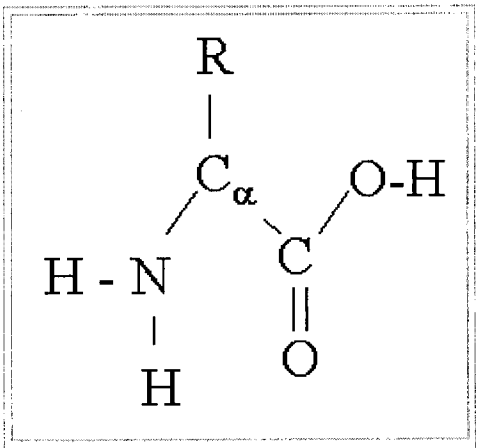
The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. These peptide bonds provide rigidity to the protein. The two ends of the amino acid chain are referred to as the C-terminal end or carboxyl terminus (C-terminus) and the N-terminal end or amino terminus (N-terminus) based on the nature of the free group on each extremity.

The various types of secondary structure are defined by their patterns of hydrogen bonds between the main-chain peptide groups. However, these hydrogen bonds are generally not stable by themselves, since the water-amide hydrogen bond is generally more favorable than the amide-amide hydrogen bond. Thus, secondary structure is stable only when the local concentration of water is sufficiently low, e.g., in the molten globule or fully folded states.

Similarly, the formation of molten globules and tertiary structure is driven mainly by structurally *non-specific* interactions, such as the rough propensities of the amino acids and hydrophobic interactions. However, the tertiary structure is *fixed* only when the parts of a protein domain are locked into place by structurally *specific* interactions, such as ionic interactions (salt bridges), hydrogen bonds and the tight packing of side chains. The tertiary structure of extracellular proteins can also be stabilized by disulfide bonds, which reduce the entropy of the unfolded state; disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment.

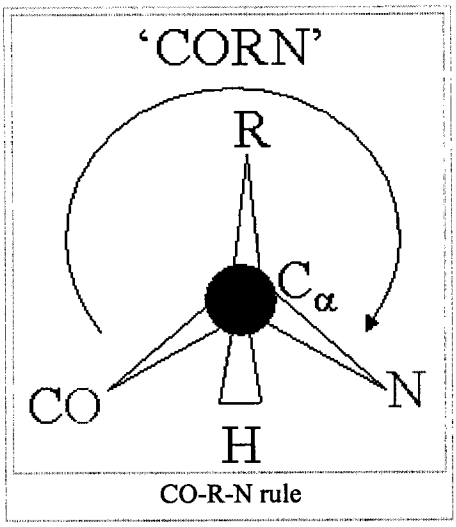
Structure of the amino acids

An α-amino acid consists of a part that is present in all the amino acid types, and a side chain that is unique to each type of residue. The C_α atom is bound to 4 different molecules (the H is omitted in the diagram); an amino group, a carboxyl group, a hydrogen and a side chain, specific for this type of amino acid. An exception from this rule is proline, where the hydrogen atom is replaced by a bond to the side chain. Because the carbon atom is bound to four different groups it is chiral, however only one of the isomers occur in biological proteins. Glycine however, is not chiral since its side chain is a hydrogen atom. A simple mnemonic for correct L-form is "CORN": when the C_α atom is viewed with the H in front, the residues read "CO-R-N" in a clockwise direction.



An α-amino acid

The side chain determines the chemical properties of the α-amino acid and may be any one of the 20 different side chains:

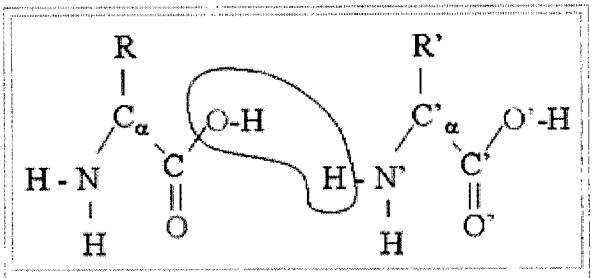


CO-R-N rule

Proline	PRO					90	
Tryptophan	TRP	W	1.0	186		163	P
Tyrosine	TYR	Y	2.2	163	10.1	141	P
Valine	VAL	V	6.0	99		105	H

The 20 naturally occurring amino acids can be divided into several groups based on their chemical properties. Important factors are charge, hydrophobicity/hydrophilicity, size and functional groups. The nature of the interaction of the different side chains with the aqueous environment plays a major role in molding protein structure. Hydrophobic side chains tends to be buried in the middle of the protein, whereas hydrophilic side chains are exposed to the solvent. Examples of hydrophobic residues are: Leucine, isoleucine, phenylalanine, and valine, and to a lesser extent tyrosine, alanine and tryptophan. The charge of the side chains plays an important role in protein structures, since ion bonding can stabilize proteins structures, and an unpaired charge in the middle of a protein can disrupt structures. Charged residues are strongly hydrophilic, and are usually found on the out side of proteins. Positively charged side chains are found in lysine and arginine, and in some cases in histidine. Negative charges are found in glutamate and aspartate. The rest of the amino acids have smaller generally hydrophilic side chains with various functional groups. Serine and threonine have hydroxylgroups, and aspargine and glutamine have amide groups. Some amino acids have special properties such as cysteine, that can form covalent disulfide bonds to other cysteines, proline that is cyclical, and glycine that is small, and more flexible than the other amino acids.

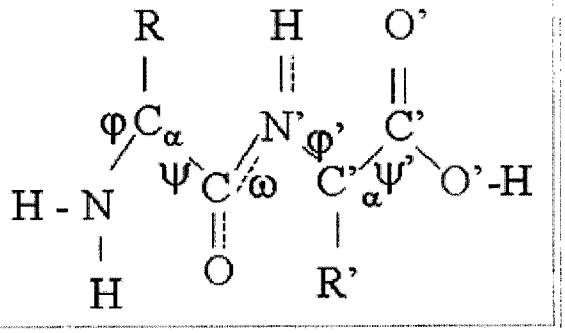
The peptide bond



Two amino acids

Two amino acids can be combined in a condensation reaction. By repeating this reaction, long chains of residues (amino acids in a peptide bond) can be generated. This reaction is catalysed by the ribosome in a process known as translation. The peptide bond is in fact planar due to the delocalization of the electrons from the double bond. The rigid peptide dihedral angle, ω (the bond between C_1 and N) is always close to 180 degrees. The dihedral angles φ (the bond between N and Cα) and psi ψ (the bond between Cα and C_1) can have a certain range of possible values. These angles are the degrees of freedom of a protein, they control the protein's three dimensional structure. They are restrained by geometry to allowed ranges typical for particular secondary structure elements, and represented in a

Ramachandran plot. A few important bond lengths are given in the table below.



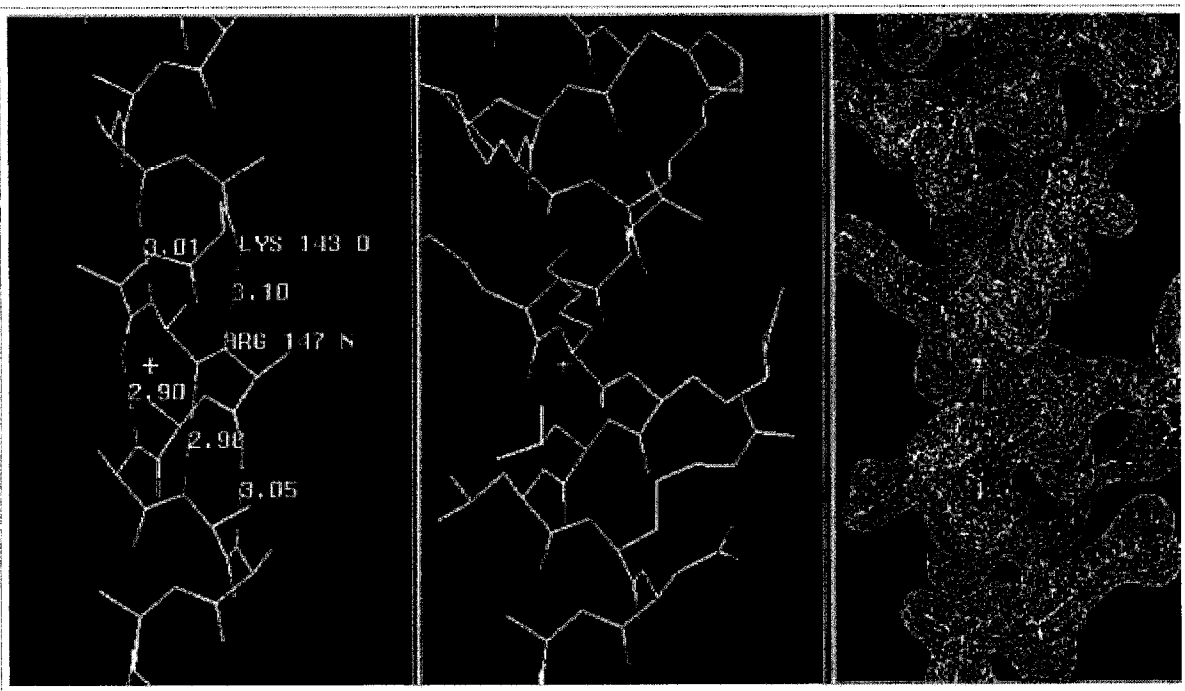
Bond angles for ψ and ω

Primary structure

The sequence of the different amino acids is called the primary structure of the peptide or protein. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code. Post-transcriptional modifications such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene.

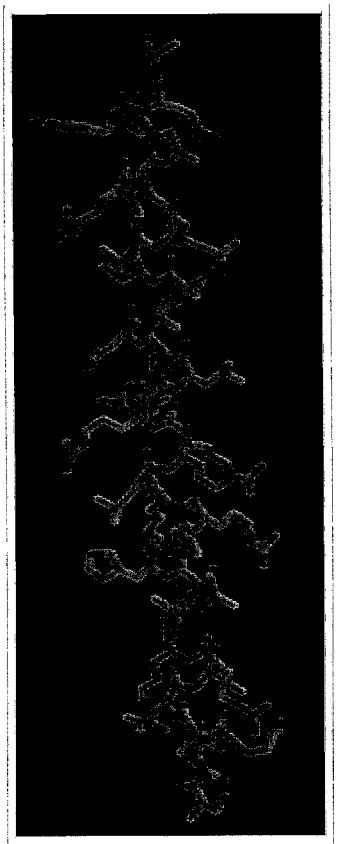
Secondary structure

By building models of peptides using known information about bond lengths and angles, the first elements of secondary structure, the alpha helix and the beta sheet, were suggested in 1951 by Linus Pauling and coworkers.[1] Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. These secondary structure elements only depend on properties that all the residues have in common, explaining why they occur frequently in most proteins. Since then other elements of secondary structure have been discovered such as various loops and other forms of helices. The part of the backbone that is not in a regular secondary structure is said to be random coil. Each of these two secondary structure elements have a regular geometry, meaning they are constrained to specific values of the dihedral angles ψ and φ. Thus they can be found in a specific region of the Ramachandran plot.

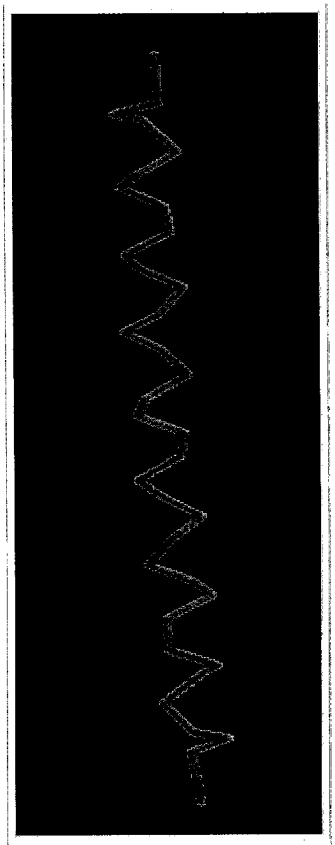


The left panel shows the hydrogen bonding in an actual α-helix backbone. Note that the *n*th residue O (Lys 1 bonds to the (*n*+4)th following residue's N (Arg 147). The actual values of some displayed H-bond distances you some idea about the variations to expect within a helix. The center panel includes the side chains which omitted in the left panel for clarity. You see the side chains pointing towards the N-terminal of the chain (lo residue numbers) and thus it is usually possible to determine the direction of the helix quite well during init model building. A 0.2 nm electron density is shown in the right panel

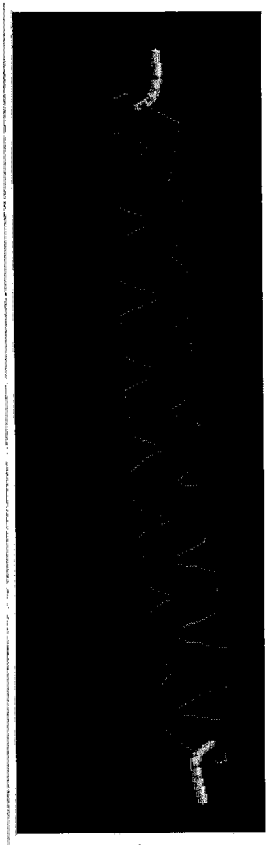
Here are some more representation of the same helix.



Ball and stick model



Backbone



Secondary structur
cartoon ("ribbon" o
"linguini diagram"

The hydrogen bond network in a 2-stranded, antiparallel β-sheet. The side chains are sticking out above or below the plane of the picture. It less clear cut than in the case of the helix, in which direction to initially trace a beta sheet strand. The beta sheet can be infinitely extended due to the repeatable H-bonding pattern to either side of a strand.

Turns, loops and a few other secondary structure elements such as a 3-10 helix complete the picture. We have now enough pieces to assemble a complete protein, displaying its typical tertiary structure.

Tertiary structure

The elements of secondary structure are usually folded into a compact shape using a variety of loops and turns. The formation of tertiary structure is usually driven by the burial of hydrophobic residues, but other interactions such as hydrogen bonding, ionic interactions and disulfide bonds

can also stabilize the tertiary structure. The tertiary structure encompasses all the noncovalent interactions that are not considered secondary structure, and is what defines the overall fold of the protein, and is usually indispensable for the function of the protein.

Quaternary structure

The quaternary structure is the interaction between several chains of peptide bonds. The individual chains are called subunits. The individual subunits are not necessarily covalently connected, but might be connected by a disulfide bond. Not all proteins have quaternary structure, since they might be functional as monomers. The quaternary structure is stabilized by the same range of interactions as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, and a tetramer if it contains four subunits. Multimers made up of identical subunits may be referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits may be referred to with a prefix of "hetero-" (e.g. a heterodimer). Tertiary structures vary greatly from one protein to another. They are held together by glycosydic and covalent bonds.

Side chain conformation

The atoms along the side chain are named with Greek letters in Greek alphabetical order: α, β, γ, δ, ϵ and so on. C_α refers to the carbon atom closest to the carbonyl group of that amino acid, C_β the second closest and so on. The C_α is usually considered a part of the backbone. The dihedral angles around the bonds between these atoms are named $\chi 1$, $\chi 2$, $\chi 3$ etc. E.g. the first and second carbon atom in the side chain of lysine is named α and β, and the dihedral angle around the α-β bond is named $\chi 1$. Side chains can be in different conformations called gauche(-), trans and gauche(+). Side chains generally tend to try to come into a staggered conformation around $\chi 2$, driven by the minimization of the overlap between the electron orbitals of the hydrogen atoms.

Domains, motifs, and folds in protein structure

Many proteins are organized into several units. A structural domain is an element of the proteins overall structure that is self-stabilizing and often folds independently of the rest of the protein chain. Many domains are not unique to the protein products of one gene or one gene family but instead appear in a variety of proteins. Domains often are named and singled out because they figure prominently in the biological function of the protein they belong to; for example, the "calcium-binding domain of calmodulin". Because they are self-stabilizing, domains can be "swapped" by genetic engineering between one protein and another to make chimeras. A motif in this sense refers to a small specific combination of secondary structural elements (such as helix-turn-helix). These elements are often called supersecondary structures. Fold refers to a global type of arrangement, like helix-bundle or beta-barrel. Structure motifs usually consist of just a few elements, e.g. the 'helix-turn-helix' has just three. Note that while the *spatial sequence* of elements is the same in all instances of a motif, they may be encoded in any order within the underlying gene. Protein structural motifs often include loops of variable length and unspecified

structure, which in effect create the "slack" necessary to bring together in space two elements that are not encoded by immediately adjacent DNA sequences in a gene. Note also that even when two genes encode secondary structural elements of a motif in the same order, nevertheless they may specify somewhat different sequences of amino acids. This is true not only because of the complicated relationship between tertiary and primary structure, but because the size of the elements varies from one protein and the next. Despite the fact that there are about 100,000 different proteins expressed in eukaryotic systems, there are much fewer different domains, structural motifs and folds. This is partly a consequence of evolution, since genes or parts of genes can be doubled or moved around within the genome. This means that, for example, a protein domain might be moved from one protein to another thus giving the protein a new function. Because of these mechanisms pathways and mechanisms tends to be reused in several different proteins.

Protein folding

The process by which the higher structures form is called protein folding and is a consequence of the primary structure. A unique polypeptide may have more than one stable folded conformation, which could have a different biological activity, but usually, only one conformation is considered to be the active, or native conformation.

Structure classification

Several methods have been developed for the structural classification of proteins. These seek to classify the data in the Protein Data Bank in a structured order. Several databases exist which classify proteins using different methods. SCOP, CATH and FSSP are the largest ones. The methods used are purely manual, manual and automated, and purely automated. Work is being done to better integrate the current data. The classification is consistent between SCOP, CATH and FSSP for the majority of proteins which have been classified, but there are still some differences and inconsistencies.

Protein structure determination

Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows one to measure the 3D density distribution of electrons in the protein (in the crystallized state) and thereby infer the 3D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by Nuclear Magnetic Resonance techniques, which can also be used to determine secondary structure. Note that aspects of the secondary structure as whole can be determined via other biochemical techniques such as circular dichroism. Secondary structure can also be predicted with a high degree of accuracy (see next section). Cryo-electron microscopy has recently become a means of determining protein structures to high resolution (less than 5 angstroms or 0.5 nanometer) and is anticipated to increase in power as a tool for high resolution work in the next decade. This technique is still a valuable resource for researchers working with very large protein complexes such as virus coat proteins and amyloid fibers.

A rough guide to the resolution of protein structures

Resolution	Meaning
>4.0	Individual coordinates meaningless
3.0 - 4.0	Fold possibly correct, but errors are very likely. Many sidechains placed with wrong rotamer.
2.5 - 3.0	Fold likely correct except that some surface loops might be mismodelled. Several long, thin sidechains (lys, glu, gln, etc) and small sidechains (ser, val, thr, etc) likely to have wrong rotamers.
2.0 - 2.5	As 2.5 - 3.0, but number of sidechains in wrong rotamer is considerably less. Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small. Water molecules and small ligands become visible.
1.5 - 2.0	Few residues have wrong rotamer. Many small errors can normally be detected. Folds are extremely rarely incorrect, even in surface loops.
0.5 - 1.5	In general, structures have almost no errors at this resolution. Rotamer libraries and geometry studies are made from these structures.

Computational prediction of protein structure

The generation of a protein sequence is much simpler than the generation of a protein structure. However, the structure of a protein gives much more insight in the function of the protein than its sequence. Therefore, a number of methods for the computational prediction of protein structure from its sequence have been proposed. *Ab initio* prediction methods use just the sequence of the protein. Threading uses existing protein structures.

Rosetta@home is a distributed computing project which tries to predict the structures of proteins with massive sampling on thousands of home computers.

Software

There are many available software packages, such as free web-based STING, used to visualize and analyze protein structures. Another example is the FeatureMap3D (http://www.cbs.dtu.dk/services/FeatureMap3D/) web-server which can visualize the quality of a protein-protein alignment in 3D and be used to map *sequence feature annotation* such as the underlying Intron/Exon structure onto a protein structure.

Several packages, such as Quantum Pharmaceuticals software[2], can be used to predict conformational changes of proteins and its influence on protein's functions.

Several methods have been developed to compare structures of different proteins. Please see structural alignment.

Computational tools are also frequently employed to check experimental and theoretical models of protein structures for errors (examples: ProSA (http://www.came.sbg.ac.at/typo3/index.php?

id=prosa), NQ-Flipper (https://flipper.services.came.sbg.ac.at/), Verify3D (http://www.doe-mbi.ucla.edu/Services/Verify_3D/), ANOLEA (http://www.swissmodel.unibas.ch/anolea/), WHAT_CHECK (http://swift.cmbi.ru.nl/gv/whatcheck/)).

References

1. ^ PAULING L, COREY RB, BRANSON HR. Proc Natl Acad Sci U S A. 1951 Apr;37(4):205-11. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. PMID 14816373 (http://www.ncbi.nlm.nih.gov/pubmed/14816373)
2. ^ Quantum Pharmaceuticals software (http://www.q-pharm.com/)

Further reading

- Habeck M, Nilges M, Rieping W (2005). "Bayesian inference applied to macromolecular structure determination (http://www.spineurope.org/publications/Habeck%20et%20al%20031912%202005.pdf)". *Physical review. E, Statistical, nonlinear, and soft matter physics* 72 (3 Pt 1): 031912. PMID 16241487 (http://www.ncbi.nlm.nih.gov/pubmed/16241487). (Bayesian computational methods for the structure determination from NMR data)

External links

- ProSA-web (https://prosa.services.came.sbg.ac.at/prosa.php) Web service for the recognition of errors in experimentally or theoretically determined protein structures
- NQ-Flipper (https://flipper.services.came.sbg.ac.at/) Check for unfavorable rotamers of Asn and Gln residues in protein structures
- servers (http://swift.cmbi.ru.nl/) That check nearly 200 aspects of protein structure, like packing, geometry, unfavourable rotamers in general of for Asn, Gln and His especially, strange water molecules, backbone conformations, atom nomenclature, symmetry parameters, etc.
- Bioinformatics course (http://swift.cmbi.ru.nl/teach/B1/). An interactive, fully free, course explaining many of the aspects discussed in this wiki entry.

Retrieved from "http://en.wikipedia.org/wiki/Protein_structure"
Categories: Protein structure
Hidden categories: All articles with unsourced statements | Articles with unsourced statements since February 2008

- This page was last modified on 7 March 2008, at 07:33.
- All text is available under the terms of the GNU Free Documentation License. (See **Copyrights** for details.) Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a U.S. registered 501(c)(3) tax-deductible nonprofit charity.